PRIORITIES IN THE PURSUIT OF PREPARING A GENERATIVE LEXICON FOR TAMIL

S. Rajendran
Tamil University
Thanjavur 613 005
Email: raj_ushush@yahoo.com

1. Introduction

The development of formal frameworks for computational lexical semantics should respond to two needs: capturing the richness of language as revealed in both meaning variation and systematic polysemy, and providing a viable and testable model for building large-scale lexical resources for natural language processing. The generative lexicon theory propounded by Pustejovsky (1994, 1995, 2001) presents a novel and exiting theory of lexical semantics that addresses the problem of the 'multiplicity of word meaning'–that is, how we are able to give an infinite number of senses to words with finite means. It is proposed here to explore the primary steps to be taken and priorities involved in building a generative lexicon for Tamil using the already available resources.

2. Generative Lexicon Theory

The generative lexicon (shortly GL) presents a novel and exciting theory of lexical semantics that addresses the problem of the "multiplicity of word meaning"- that is, how we are able to give an infinite number of senses to words with finite means. As the first formally elaborated theory of generative approach to word meaning, it lays the foundation for an implemented computational treatment of word meaning that connects explicitly to a compositional semantics. In contrast to static view of word meaning (where each is word is characterized by a predetermined number of word senses) that imposes a tremendous bottleneck on the performance capability of any natural language processing, Pustejovsky proposes that the lexicon becomes an active and central component in the linguistic description. The essence of his theory is that the lexicon functions generatively, first by providing a rich and expressive vocabulary for characterizing lexical information; then by developing a frame work for manipulating fine-grained distinctions in word descriptions; and finally, by formalizing a set of mechanisms for specialized composition of aspects of such description of words, as they occur in context, extended and novel senses are generated.

According to Pustejovsky (1995:5, 2001:55) Generative Lexicon Theory is concerned with the following problems:

1. Explaining the polymorphic nature of language
2. Characterizing the semanticality of natural language utterances
3. Capturing the creative use of words in novel contexts
4. Developing a richer, co-compositional semantic representation

Pustejovsky (1995:61) characterize a generative lexicon as a computational system involving at least the following levels of representation:

1. ARGUMENT STRUCTURE: Specification of number and type of logical arguments
2. EVENT STRUCUTE: Definition of the event type of an expression and its subeventual structure
3. QUALIA STRUCTURE: A structural differentiation of the predicative force for a lexical item
4. LEXICAL INHERITANCE STRUCTURE: Identification of how a lexical structure is related to other structures in this type of lattice

Pustejovsky (1995:76, 2001:56) assumes that word meaning is structured on the basis of four generative factors, or qualia roles, that capture how humans understand objects and relations in the world and provide the minimal explanation for the linguistic behaviour of lexical items.

CONSTITUTIVE: the relation between an object and its constituent parts
FORMAL: the basic category that distinguishes the object within a larger domain
TELIC: the object's purpose and function
AGENTIVE: factors involved in the object's origin or "coming into being."

A set of generative devices connects these four levels, providing for the compositional interpretation of words in context. Included in these generative operations are the following semantic transformations, all involving well-formedness conditions on type combinations:

TYPE COERCION: where a lexical item or phrase is coerced to a semantic interpretation by a governing item in the phrase, without change of its syntactic type.
SELECTIVE BINDING: where a lexical item or phrase operates specifically on the substructure of a phrase, without changing the overall type in the composition.
CO-COMPOSITION: where multiple elements within a phrase behave as functors, generating new non-lexicalized sense for the words in composition. This also includes cases of underspecified semantic forms becoming contextually enriched, such as manner co-composition, feature transcription, and light verb specification.

The qualia structure is the core of the generative properties of the lexicon, because it provides a general strategy for creating increasingly specific concepts with conjunctive properties. A simple schematic description of a lexical item, α, using this representation is shown below:

α
ARGSTR = ARG1=x
...
CONST = what x is made of
QUALIA = FORMAL = what x is

TELIC = function of x
AGENTIVE = how x come into being

The lexical structure for *book* as an object can then be represented as follows:

book

ARGSTR = ARG1 = y:information
ARG2 = x:phys_obj. information.phy_obj
FORM = holds (x,y)

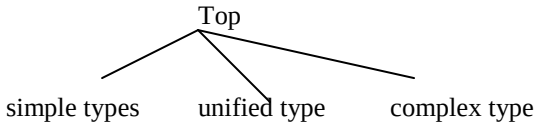
QUALIA = TELIC = read (e,w,x.y)
AGENT = write (e', v, x, y)

Pustejovsky (1995) defines the semantics of a lexical item as a structure involving different components. One of these is qualia structure, which is rich and structured representation of the rational force of a lexical item. What is peculiar about GL is that qualia permits a much richer description of meaning than either a simple decompositional view or a purely relational approach to word meaning allow. That is, it expresses different/orthogonal aspects of word meaning instead of a one-dimensional inheritance (even multiple), which can only capture standard hyponymy/hypernymy relations. The adequacy of qualia relations for capturing key aspect of word meaning becomes apparent when consulting dictionary definitions. The elements of meaning easily map on dimension(s) expressed via qualia roles. Furthermore, these relations become particularly crucial for those sense definitions which have an underspecified genus term. Not all dimension of meaning are always explicitly expressed in the definition of a lexical item. Some of them are inherited by virtue of its membership to a semantic type. Although qualia relations easily emerge form dictionary definition, the formal expression of a specific value for a quale is sometimes quite problematic.

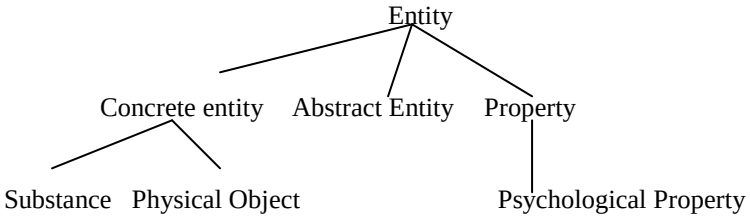
Busa el al in their paper (2001:333-349) presents the recent extension of Generative Lexicon theory (Pustejovsky, 1995, 1998). According to them the development of lexical resources should be guided by an underlying frame work for structuring word meaning and generating concepts, which satisfies both ontological considerations as well as the need to capture linguistic generalizations. Explicitly defined ontologies play a very important role in lexicon design: the meaning of an item is partly determined by the position in the ontology occupied by the concept or concepts it expresses; they allow the lexicon designer to have a clear and comprehensive view of the overall relations between the lexical items in the knowledge base. The general principles underlying the design of most current ontologies are based on two main assumptions. The first is that the ontology is taken to be a conceptualization of the world as we understand it; as a result elements of the ontology are identified in terms of language independent criteria. The second is that the links between nodes determining the major conceptual partitions often reduce to ISA relation only.

Lexical items with different degrees of complexity, as determined by their internal semantic constituency, i.e., qualia structure, are organized in the lexicon in terms of a tripartite structure distinguishing

between simple type, unified types, and complex types at the top level (cf. Pustejovky, 1998). This can be shown schematically as given below:

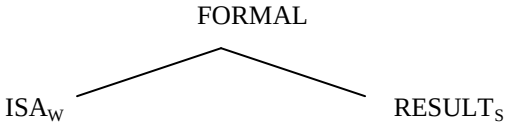


Intuitively, the type of a lexical item is simple if the item is uniquely defined in terms of a taxonomic relation to another entity in a hierarchy. Natural kind entities are such an example. Similarly, "information", "abstract," are also member of the set of simple types when they do not involve richer information concerning other aspects of meaning.



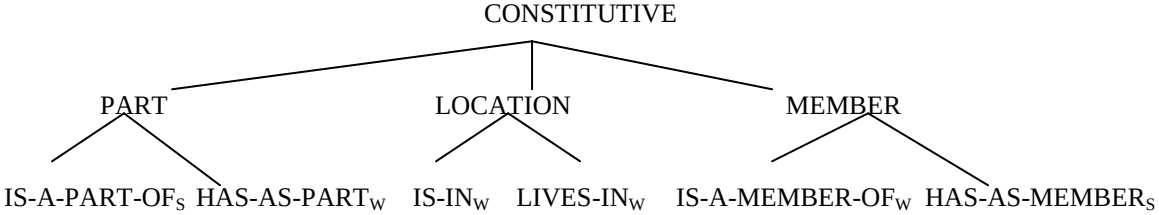
Unified types implement the principle of orthogonal inheritance (Pustejovky and Boguraev, 1993), which allow a lexical item to fall into multiple classes along different dimensions of meaning. A unified type is created by recursively combining a simple or another unified type with additional elements from qualia structure. Finally, complex types model the properties of logically polysemous lexical items.

Weak and strong qualia determine whether or not orthogonal components of meaning give rise to simple or unified types. For each role in the qualia, this property is explicitly marked by means of the subscripts S and W, as shown in the FORMAL type hierarchy bellows (Busa et al, 2001:341)



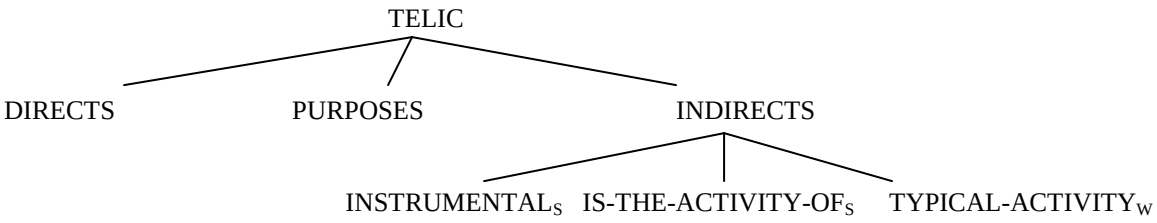
The hierarchy for FORMAL is actually quite simple and it only distinguishes between the standard ISA relation (i.e., a weak quale), and results, (i.e., a strong quale), marking the resulting state for complex event structures denoting transitions.

The qualia set for the CONSTITUTIVE role involves a number of subtypes, which expresses different constitutive relations that contribute to the semantic description of various concepts (Busa et al, 2001:341):

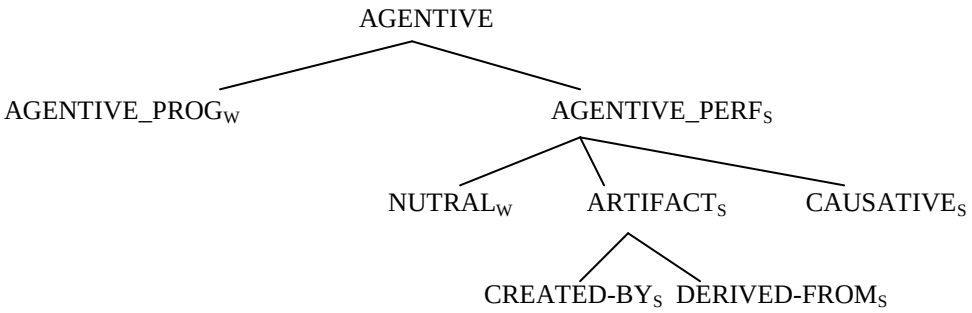


The strong CONSTITUTIVE types are in the construction of unified types such as “groups,” “organ,” “headquarter,” “component,” “ingredient.”

The TELIC involves a set of subtypes that distinguish lexical items according to a number of parameters.



Finally, the AGENTIVE qualia set distinguishes between persistent and temporary properties of the event encoded therein.



3. Nominal Semantics

Dictionary definitions of nouns are normally written in such a way that one can identify for each headword, a “genus term”, and these are related via an IS-A relation. Information following the genus term, the differentia, serves to differentiate the headword from other headwords with the same genus. Now let us examine the meanings of certain lexical items given in *kiriyaavin taRkaalattamiz akaraati* (shortly KTTA, which in machine readable form) and explore how much of information we can gather for constructing qualia structure: (The English glosses are given to facilitate understanding.)

> *puttakam* ‘book’ is defined in KTTA as follows:
> *puttakam pe*. 1. *paTippataRku eeRRavakaiyil aTTaipooTTu iNaitta acciTTa taaLkaLin tokuppu* ‘printed pages tied together by binding into a collection to enable reading.’2. *ezutuvataRku eeRRavakaiyil veRRuttaaLkaLai aTTai pooTTu iNaitta tokupp*pu ‘plain paper tied together into a collection to enable writing’.

tokuppu ‘collection’ is the genus (i.e. formal role) of *puttakam*, which can be easily identified being the head noun of the clause. *aTTaipooTTu iNaitta* and *accaTitta taaLkaL* ‘printed papers’ are attributes (constitutive role) which helps to distinguish *puttakam* ‘book’ form other printed materials, for example, *accaTitta aRivippu*

'printed notice'. *paTippataRku* 'for reading' stands to denote the function (i.e. telic role) of *puttakam*. From the definition given in KTTA we can arrive at the following qualia structure for *puttakam*.

***puttakam* 'book**

ARGSTR =	ARG1	= x: phys_obj
	CONST	= made_ of: *aTTai* 'cover', *accaTitta taaLkaL* 'printed papers'
	FORM	= *tokuppu* 'collection' (x)
QUALIA =	TELIC	= *paTi* 'read' (x,z)
	AGENT	= *accaTi* 'print' (x), *pooTTu iNai* 'bind by': *aTTai* 'cover'

The second meaning stands to denote a *notebook*. The agentive information that *puttakam* is the product of writing is not given in the meaning. Instead it is given as the product of printing. Also the fact that *puttakam* is ambiguous like *book*, denoting information and physical object cannot be had from the definition.

Let us look at the meaning of the *katti* found in KTTA:
katti 'knife' *paTTaiyaana ulookattakaTTin oorangkaLil kuurmai uTaiyataaka uLLa veTTum karuvi* 'a cutting instrument of flat meal plate with sharp edges'

In the definition *karuvi* 'instrument' is the genus term of the headword *katti* and *paTTaiyaana ulookattakaTTin oorangkaLil kuurmai uTaiyataaka uLLa* 'flat meal plate with sharp edges' and *veTTum* 'that which can cut' yields differentia. The noun phrase *paTTaiyaana ulookattakaTTin* 'in band of metal plate' gives made_of constitutive role and the clause *oorangkaLil kuurmai uTaiyataaka uLLa* 'with sharp edges' gives the has_as_ part constitutive role and the *veTTum* gives the telic role. *katti* can be represented by the following structure:

***katti* 'knife'**

ARGSTR =	ARG1	= x: phys_obj
	CONSTIT	= made_of: *paTTaiyaana ulookattakaTu* 'band of metal plate'.
QUALIA =		has_as_part: *kuurmaiyaana oorangkaL* 'sharp edges'
	FORMAL	= *karuvi* 'instrument' (x)
	TELIC	= *veTTu* 'cut' (x,z)
	AGENTIVE	= ?

That agentive role that *katti* is an artifact is not explicitly spelt in the definition, but it can be inferred form the genus term *karuvi* 'instrument'.

4. Verbal Semantics

The qualia structure associated with causative predicates (i.e. TRANSACTIONS) or verbs are typically analyzed as involving an initial act or process followed by a resulting state. These two phases map directly into the AGENTIVE and FORMAL qualia roles, respectively. The transitive form of the verb *uTai* '*break* is illustrated below (Pustejovsky (1995:80):

***uTai* 'break'**

E1 = e1: *ceyalpaangku* 'process'

EVENTSTR= E2 = e2: *ndilai* 'state'

RESTR = <α

FORMAL = *uTai* 'broken' (e2,y)

QUALIA= AGENTIVE = *uTai_ceyal* (e1,x,y)

The qualia structure of *ooTu* 'run' and *tuungku* 'sleep' are differentiated by the AGENTIVE and FORMAL roles.

***tuungku* 'run'**

EVENTSTR = E1 = e1: *ceyaRpaangku* 'process'

QUALIA = AGENTIVE = *ooTu_ceyal* (e1,x)

***tuungku* 'sleep'**

EVENTSTR = E1 = e1: *ceyalpaangku* 'process'

QUALIA = FORMAL = *tuungku* (e1, x)

Now let us examine definitions of a few verbs and how far the information fulfill qualia structure. The verb *tangku* 'stay' is described in KTTA as follows:

tangku 'stay' 1: (*kuRippiTTa ndookkattukkaaka oor iTattukkuc cenRu) kuRippiTTa kaalam varai iruttal* 'to go and be at a place for a fixed period of time with a purpose'

In line with Busa et al (2001) we can represent the lexical structure of *tangky* 'stay' in the following fashion:

***tangku* 'stay'**

type: simple:location_formal$_{isa}$

argstr: arg1: x: living_entity

arg2: y: location

evenstr:[e1:e1:*ndilai* 'state']

qualia: [formal:at (e1,x,y)]

The definition in KTTA emphasizes being at a place for certain period. The lexical structure given by Busa et al does not account for time, only location is accounted. The information needed for lexical structure can be had form the definition and the illustrative sentences. KTTA defines *anuppu* 'send' in the following fashion:

anuppu 'send'1: (*oruvarai*) *oor iTattukkup pookacceytal* 'to make a person go to a place'; (*onRai*) *oor iTam cenRaTaiyac ceytal* 'causing a thing to reach a place'

***anuppu* 'send'**

type: unifed:location_formal$_{result}$_agentive$_{cause}$_telic$_{purp}$

argustr: arg1 : x:living_entity
arg2 : y:entity
arg3 : z:entity

evenstr: e1: e1:state
e2: e2:process
rest: e1 precedes e2

qualia: formal (¬at (e1,x,y)
telic: at (e1,z,y)
agentive: act (e2,x,y)

5. Polysemy

A proper treatment of the description of the semantics of lexical items in the language should permit us to adequately describe the behaviour of ambiguity as well as the lexical process of selection process in the grammar. Weinreich (1964) distinguishes two types of ambiguity, the first of which he calls contrastive ambiguity. This is seen where a lexical item accidentally carries two distinct and unrelated meanings (i.e. homonymy). For example, the lexical item *bank* found in the following sentences exhibit contrastive senses.

koovilukkuc cella nduuRu paTikaL eeRaveeNTum
'One has to climb hundred steps to reach the temple'
avan pattu paTi arici vaangkinaan
'He bought ten measures of rice'

Complementary ambiguity exhibits a different type of relation between senses. The sense alternation is just one of many alternations that can be described as logical polysemies, where the noun seems to have systematically related senses.

Count/Mass alternations; *aaTu* 'goat'
a. *aaTu vayalil ooTukinRatu* 'The goat is running in the field.
b. *kaNNan acaiva uNavakattil aaTu cappiTTaan*
'Kannan ate mutton in the non-vegetarian hotel'
Container/Containee alternations*;* bottle
a. *raatai kuppiyai uTaittaaL* 'Radha broke the bottle.'
b. *kuzandtai kuppiyai kaali ceytatu* 'The baby finished the bottle.'

The most direct way to account for the polysmies described above is to allow the lexicon to have multiple listing of words, each annotated with a separate meaning or lexical sense. This is certainly the simplest means of

encoding sense variation in a lexical item. Pustejovsky (1995:34) calls such a dictionary as a Sense Enumeration Lexicon (SEL), and characterize it is directly as follows:

> A Lexicon L is a Sense Enumeration Lexicon if and only if for every word w in L, having multiple senses s1,...,sn, associated with the word, then the lexical entries expressing these senses are sorted as {ws1,...,wsn}.

SEL represent the contrastive senses of *paTi* in the following fashion:

***paTi* 1 'step'**	***paTi* 2 'a kind of measure'**
CAT = count_noun	CAT = count_noun
GENUS = *paTikkaTTin pakuti* 'part of steps'	GENUS = *aLavai* 'measure'

Also SEL represents the complementary senses of *aaTu* in the following fashion.

***aaTu* 1 'goat'**	***aaTu* 2 'mutton'**
CAT = count_noun	CAT= mass_noun
GENUS = *vilangku* 'animal'	GENUS = *iRaicci* 'meat'

Pustejovsky (1995: 38) restate the SEL's account of contrastive and complementary senses as follows:

> A Lexicon L is a Sense Enumeration Lexicon if and only if for every word w in L, having multiple senses $s_1,...,s_n$ associate with that word, then:
>
> (i) if $s_1,...,s_n$ are contrastive senses, the lexical entries expressing these senses are stored as $w_{s1},...w_{sn.}$
> (ii) if $s_1,...s_n$ are complementary senses, the lexical entry expressing these senses is stored as w{s1,...sn}.
> Every ambiguity is either represented by (i) or (ii) above.

Though Pustejovsky points out the advantage of this model of lexical description, he also points out that the SEL model is inadequate for the purpose of linguistic theory. But unfortunately KTTA does not stand to fulfill the requirement of a SEL model completely. For example, KTTA does not enumerate the mass-noun sense of the lexical item *koozi* 'fowl', though it is a possible sense.

> ***koozi*** *pe.* (*potuvaaka*) *viiTu, paNNai aakiyavaRRil muTTaikkaakavum iRaiccikkaakavum vaLarkkappaTum atika uyaram paRakkaata p*aRavai 'a type of bird which cannot fly high and which are generally reared in houses, farms, etc. for the sake of eggs and meat'

The lexical item *jannal* 'window' is polysemous expressing aperture sense and physical object sense which can be represented as follows:

jannal '**window1'**	*jannal* '**window2'**
CAT = count_noun	CAT = count_noun
GENUS = aperture	GENUS = physical_obj

KTTA does not account for the the physical object sense of *jannal* 'window'.

> ***jannal*** *pe. kaTTiTam poonRavaRRiRil kaaRRum veLiccamum varuvataRku eeRRa vakaiyil tiRandtu muuTakkuuTiyataaka cuvaril amaikkappaTTirukkum tiRappu*. 'an opening in the wall of a building which can be closed or opened with the facility of the light and wind to come inside'

Pustejovsky (1995:39-60) points out the limitations of SELs. KTTA also shows all these limitations being a dictionary and not a generative lexicon.

6. Conclusion

GL argues that compositional meaning takes place under predictable circumstances, provided there is an adequate set of principles for describing lexical structure. Senses need not be enumerated but can be derived generatively form ricer representations of words and compositional mechanisms of combining them. The nature of individual lexical representations characterizes literal meaning, which involves two closely tied elements: analytic and conventionalized knowledge. The analytic component is the structuring itself, namely, qualia structue; the conventionalized component is represented by the value that fills each qualia role.

Reference

Bouillon, P and Busa, F. 2001 "Qualia and the Structuring of Verb Meaning." In P. Bouillon and F. Busa (eds.). The Language of Word Meaning. Cambridge: Cambridge University Press, 149-167.

Busa, F & Calzolari, N and Lenci, A. 2001. "Generative Lexicon and the SIMPLE Model: Developing Semantic Resources for NLP." In F. Busa and P. Boulillon (eds). Language of Word Meaning. Cambridge: Cambridge University Press, 333-349.

Climent,

Foddor, J.A. and Lepore, E. 2001. " The Emptiness of the Lexicon: Critical Reflections on J. Pustejovky's "The Generative Lexicon"." In Busa and P. Boulillon (eds). Language of Word Meaning. Cambridge: Cambridge University Press, 28-49.

McGilvary, J. 2001. Chomsky on the Creative Aspect of Language Use and Its Implications for Lexical Semantic Studies." In F. Busa and P. Boulillon (eds). Language of Word Meaning. Cambridge: Cambridge University Press, 5-27.

Pustejovksy, J. 1994. “A Richer Characterization of Dictionary Entries: The Role of Knowledge Representation.” In B.T.S. Atkins, and A. Zampolli (eds.). Computational Approach to Lexicon. Oxford: Oxford University Press.

-------1995a. “Linguistic Constraint on Type Coercion.” P. Saint-Dizoer and E. Viegas (eds.). Computational Lexical Semantics. Cambridge: Cambridge University Press.

-------1995b. The Generative Lexicon. Cambridge, Massachusetts: The MIT Press.

-------2001. “Generativity and Explanation in Semantics: A Reply to Fodor and Lepore.” In

P. Bouillon and F. Busa (eds). The Language of Word Meaning. Cambridge: Cambridge University Press, 51-74.

-------2001. “Type Construction and the Logic of Concepts.” In F. Busa, and P. Boulillon (eds). Language of Word Meaning. Cambridge: Cambridge University Press, 91-123.

Pustejovky, J. and Boguraev, B. 1993. “Lexical Knowledge Representation and Natural Language Processing.” In F. Pereira and B. Grosz (eds.), Natural Language Processing. Cambridge, Mass.: MIT Press.

Rajendran, S. 1978. Syntax and Semantics of Tamil Verbs (manuscript). Ph.D. Thesis. Poona: University of Poona.

------1983. Semantics of Tamil Vocabulary. (Report of the UGC sponsored Postdoctoral Work in manuscript). Poona: Deccan College Post Graduate and Research Institute.

------ 2002. “Preliminaries to the Preparation of a Word Net for Tamil.” Language in India 2:1, www.langugeinindia.com

------ 2003. “Creating Generative Lexicon from Dictionaries: Tamil Experience.” In Recent Advances in Natural Language Processing. Myore: CIIL, 83-91.

Rajendran, S., Arulmozi S., Kumara Shanmugam, B., Baskaran, S. and Thiyagarajan, S. “Tamil WordNet.” In Proceedings of the First International Global WordNet Conference. Mysore: CIIL, 271-274.

Ruimy, N. Gola, E. and Monachini, M. 2001. Lexicography Informs Lexical Semantics: The SIMPLE Experience. In F. Busa, and P. Boulillon (eds). Language of Word Meaning. Cambridge: Cambridge University Press.

Saint-Dizoer, P and Viegas, E. (eds.) 1995. Computational Lexical Semantics. Cambridge: Cambridge University Press.

Wilks, Y.A., Slator, B.M. and Guthrie, L.M. 1996. Electric Words: Dictionaries, Computers, and Meanings. Cambridge, Mass.: MIT Press.